NAMING PERSONS IN NEWS VIDEO WITH LABEL PROPAGATION

Phi The Pham, Marie-Francine Moens

Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A
B-3001 Heverlee, Belgium

Tinne Tuytelaars

Katholieke Universiteit Leuven
ESAT - PSI
Kasteelpark Arenberg 10
B-3001 Heverlee, Belgium

ABSTRACT

Labeling persons appearing in video frames with names detected from the video transcript helps improving the video content identification and search task. We develop a face naming method that learns from labeled and unlabeled examples using iterative label propagation in a graph of connected faces or name-face pairs. The advantage of this method is that it can use very few labeled data points and incorporate the unlabeled data points during the learning process. Anchor detection and metric learning for face classification techniques are incorporated into the label propagation process to help boosting the face naming performance. On BBC News videos, the label propagation algorithm yields better results than a Support Vector Machine classifier trained on the same labeled data.

***Keywords*—** Cross-media mining, image annotation.

1. INTRODUCTION

The aim of our work is to label faces present in video frames with people names extracted from video transcripts. In this research, we focus on label propagation, i.e., once we have obtained reliable name-face pair exemplars, we propagate the name to similar faces. Especially for indexation of videos, the results of our work are very useful.

More specifically, after names and faces are detected, some seed name-face pairs are sought. These can be manually selected, or automatically selected relying on some confident alignments. A label propagation algorithm is then applied on the remaining faces of the video, that maximizes the likelihood of all alignments of names and faces. In this framework it is also possible that for a given face, there is no corresponding name (so the face refers to a null name)[1]. The problem is challenging because there are often many faces in one image frame or set of frames, and many names in the transcripts of the video. Another difficulty is that the time at which people are named and the time at which the corresponding faces are shown in the image can diverge. In addition, the names in the audio stream of the video are not always correctly recognized, which forms an additional source of errors in the naming process.

We apply our face naming methods on BBC news broadcasts. Often the anchor person is shown and is named in the beginning of the broadcast. The name-face pairs corresponding to the anchor can be separately identified, which might boost the accuracy of the overall alignments.

The contributions of this paper are a graph based algorithm for learning the name-face alignments jointly from labeled and unlabeled examples. We show that the algorithm is successful and better copes with noisy face detections than using a classification model trained on the labeled data. As manual labeling is expensive, in all our experiments, we use very few labeled data. In addition, we develop an unsupervised model for naming anchor persons in news.

The remainder of this paper is organized as follows. In the next section we describe related work. In section 3 we discuss the video data preprocessing, i.e., the detection and clustering of the person names in the transcripts, the face detection and tracking in the video frames and the anchor detection in news broadcasts. Section 4 describes the label propagation algorithm that computes the probability of a name for each face. Section 5 reports on a learning metric for face classification. In section 6 the experiments are described and the results are given. Section 7 concludes the paper.

2. RELATED WORK

Several researchers have looked into the problem of linking names and faces, be it with a manual initialization (e.g. [1, 2]) or fully automatic based on temporal co-occurrence of names and faces (e.g. [3, 4, 5]). The initial information of names and faces correspondence is then used as a training set for a regular face classification process [1, 4], a character specific multiple kernel classifier [5], or for a multiple instance learning process [2]. In the Name-It system [3], based on the initial information of name and face temporal co-occurrence, a heuristic method is applied to select the name-face pairs with the most significant temporal correspondence.

Based on a large news photo with caption collection, [6]

The research was financed by the EU FP6-027978 project CLASS (Cognitive-Level Annotation Using Latent Statistical Structure) and the IWT (SBO 060051) project AMASS++ (Advanced Multimedia Alignment and Structured Summarization). Phi The Pham is funded by Katholieke Universiteit Leuven's IRO scholarship.

[1] In this paper we are not interested in the inverse problem where a detected name refers to no face or to the null face.

proposes a graph based method for finding the group of most similar faces associated with a given name. They first use the queried name to collect a set of the associated faces. Then a weighted graph is built with the nodes are the associated faces and the edges are their dissimilarities. Finally, they find the densest component - the set of highly connected nodes of the graph and consider this set of nodes to correspond to the faces of the queried person. [7] solves a different problem (unsupervised object discovery), yet uses somewhat similar techniques: a graph is constructed, with all images as vertices and edge weights based on image similarity (number of feature matches). Then they use normalized cuts to partition the graph in different clusters ('objects').

It has been shown that in classification tasks learning from labeled and unlabeled examples has a benefit on classification performance, especially when few labeled training examples are available [8, 9]. We are interested in methods that satisfy our assumption that similar faces tend to have similar name labels [10, 11]. In this paper we target such an approach for the name-face alignments.

The problem of anchor person detection is also well studied [12]. The common assumption is that, the anchor persons appear many times throughout the news broadcast and they appear in front of the same studio settings. Hence, the background images where people appear are clustered based on their color similarity. The clusters with more than one member and with the total time span higher than a threshold are considered referring to anchor persons. Other information such as audio and person faces (the presence of the anchor person corresponds to his/her speech and face) are also used to boost the anchor detection accuracy. In this paper we accurately identify the anchor persons with their name.

3. VIDEO DATA PREPROCESSING

Our methods presume video frames and related text transcripts.

3.1. Preprocessing of the transcript

3.1.1. Detection of person names

A first step is to recognize person names in the transcript. We use a named entity recognizer which is based on a maximum entropy classifier from the OpenNLP package[2], which we augment with a gazetteer of names which are extracted from the Wikipedia[3] website.

3.1.2. Clustering of the person names

In one segment of the transcript several mentions (e.g., "Al Gore", "former vice president", "he") might refer to the same person and form a coreference chain. Within one segment this noun phrase coreference resolution follows the methods of the LingPipe[4] package. To group mentions of the same person across the transcript, we use a dictionary of variant names in combination with a clustering of the coreference chains of a name, where the latter allows to resolve mentions of the single word "Bush" to "George W. Bush" and not to "Laura Bush". Then coreference chains of each text are clustered with a hierarchical single link algorithm constrained by a threshold cosine similarity for cluster membership.

[2]http://opennlp.sourceforge.net/

[3]http://en.wikipedia.org/

[4]http://www.alias-i.com/lingpipe/

3.2. Preprocessing of the video frames

3.2.1. Detection and description of faces

A parallel task regards the detection and description of the faces in the video frames. This is a challenging task under uncontrolled conditions, due to the wide variability in face appearance – especially because of changes in pose, illumination conditions, facial expressions, and partial occlusions. First, faces are detected using the OpenCV implementation of [13]. Next, we detect facial features [4] and use these as initial pose estimation for a 3D morphable face model [14] that is fitted to the data. Using such a 3D morphable model allows to estimate the pose and illumination parameters and to eliminate these irrelevant sources of variability. Also partial occlusions can be overcome this way. The model returns 40 person-specific texture components and 40 person-specific shape components, which together form the face descriptors used in this work.

3.2.2. Face tracking

A typical news footage lasts on average 30 minutes and contains around 30000 detected faces. Manually labeling all these faces is extremely time-consuming and the face label propagation process has to deal with a large amount of data points. These faces arise from just a few hundred "tracks" of a particular character each in a single shot [4]. Applying a face-tracking method reduces the amount of data to process and allows us to select the best faces in each face track to help improve the quality of face comparison and classification. We use $Kanade - Lucas - Tomasi$ [15] method to track faces using similar point tracks over frames.

3.3. Anchor/reporter detection

We adopt the approach by [12] to find the anchor face tracks. To find the candidate names for the anchors, we use the transcript clues as exploited by [1]. E.g, when we hear "*I'm Alastair Yates. Have a good night*", "*Alastair Yates*" should be the name of the anchor. Finally, to match an anchor face track and a candidate name, we adapt the timing similarity approach by *Satoh et al.* [3]. If the timing similarity between an anchor face track and a candidate name is higher than a threshold (set via observations on a held-out data set), the candidate name is assigned to the anchor face track.

In this framework, the named anchors are moved from the unlabeled face set to the training set. This will reduce the ambiguity in our name label propagation process and might improve the alignment performance.

3.4. Initial labeling of name-face seed pairs

We first annotate a number of name-face seed pairs. This annotation can be done manually by randomly selecting a number of faces and assigning a name to them, if the name is mentioned in the text. As in [4] we select faces from the ground truth data.

4. NAMING FACES WITH LABEL PROPAGATION

4.1. Presentation of the problem

Given a training set of labeled faces, the problem of choosing names for the unlabeled faces can be solved with many classification methods. Since the value of the unlabeled data is proved to be valuable and the manual annotation of all the faces in the videos is overwhelming, the class of classification methods that combine the labeled and unlabeled faces into the learning process is important to our face labeling task. Among various classification methods using labeled and unlabeled data, label propagation by a random walk process approach is promising [10, 11]. We adapt the approach of [10] to formulate a framework of name label propagation over faces.

Suppose we initially can find name labels for l faces and the remaining u faces do not have name labels yet. We denote $(f_1, n_1)...(f_l, n_l)$ the set of labeled faces where $N_l = \{n_1...n_l\} \in \{1...C\}$ are the name labels. The constraints are: the number of distinct names C is known and all the distinct name labels appear in the set of labeled faces. The set of unlabeled faces can be presented as $(f_{l+1}, n_{l+1})...(f_{l+u}, n_{l+u})$ where the name labels $N_u = \{n_{l+1}...n_{l+u}\}$ are not known yet. Let F be the set of all faces $F = \{f_1...f_{l+u}\} \in R^D$ described with D features. We will use F and N_l to predict N_u.

If a group of faces are similar, they may have the same name. The level of similarity between these faces will decide the confidence that they share the same name. Generally, the labeled faces can contribute to the naming of an unlabeled face by their names with the confidence estimated by their similarities with this unlabeled face. Moreover, the unlabeled faces can also affect the labels of each other by their similarity.

4.2. Using solely face similarities

To implement the above observation, we build a fully connected graph G where the nodes are all $l + u$ faces. The weight w_{ij} of the edge between faces f_i and f_j is the similarity between them.

The one-step transition probability T_{ij} from face j to face i can be estimated from the edge weights:

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}} \quad (1)$$

All faces have probability distributions over name labels. We define a probability matrix N of size $(l + u) \times C$ where N_i is the probability distribution of name labels over face i. The one-step name label propagation between faces can be computed by performing a matrix multiplication TN.

4.3. Using name-face pairs similarities

In the type of media like videos, we notice that: (1) Faces in the same frame mostly have different names; (2) Faces in the same face track must have the same name.

To implement the new constraints for our label propagation algorithm, we define the following additional functions: $SN(n_b, n_y)$ is equal to 1 if the names n_b and n_y are the same, 0 otherwise; $SF(f_a, f_x)$ is equal to 1 if the faces f_a and f_x are in the same video frame and their tracks overlap temporally, 0 otherwise; $ST(f_a, f_x)$ is equal to 1 if the faces f_a and f_x are in the same face track, sim_{f_a,f_x} otherwise. The third function, $ST(f_a, f_x)$, makes use of the results of the face tracking process in the sense that if two faces are in the same face track, their similarity should be 1, so that the probability for them to share the same name is high.

In this model we consider all candidate name-face alignment pairs of one news broadcast, some of which are sure alignments (labeled examples) and of the other pairs we do not know if the alignment holds. We build a completely different graph $G\prime$, now with the name-face-pairs as vertices and a new transition matrix between name-face pairs. From the above constraints, we redefine the weight $w_{i,j}$ of the edge connecting the name-face pairs $i =< f_a, n_b >$ and $j =< f_x, n_y >$ as follows:

$$w_{ij} = SN(n_b, n_y) * (1 - SF(f_a, f_x)) * ST(f_a, f_x) \quad (2)$$

The one-step transition probability T_{ij} from name-face pair j to name-face pair i can be estimated from the edge weights:

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{(l+u)C} w_{kj}} \quad (3)$$

All name-face pairs now have probability distributions N over the possibility that these pairs are true or not. The one-step name label propagation between name-face pairs can again be computed by performing a matrix multiplication TN.

4.4. The label propagation algorithm

After setting up the graph G, the transition matrix T and the label matrix N, we perform the label propagation algorithm as follows:

1. All faces/name-face pairs propagate labels for one step: $N \leftarrow TN$
2. Row normalize N to maintain the label probability interpretation.
3. Clamp the labeled faces. Repeat from step 2 until convergence of N.

Step 3 is important since the names of the labeled faces are kept and via iterations, the labels are propagated through the high density (similar) face/name-face pair regions (clusters) and settle down in the low density gaps between face/name-face pair clusters.

The label propagation algorithm is proved to converge to a simple solution [10].

4.5. Face naming after the label propagation process

Matrix N contains then the label distribution for each face/name-face pair. In the experiments below we use for each face the name with highest probability or the alignment of a name-face pair with highest probability where the probability is above a threshold λ (called name assignment threshold in the experiments below) [4]. This refusal to predict strategy leaves some faces unlabeled, or in other words these faces refer to the null name.

5. LEARNING METRIC FOR FACE CLASSIFICATION

The last problem is to learn a good metric for face similarities. A simple function to estimate the similarity between two faces f_i and f_j is defined as follows:

$$sim_{f_i f_j} = exp(-d_{f_i f_j}) = exp(-\sqrt{\sum_{k=1}^{D}(f_i^k - f_j^k)^2}) \quad (4)$$

This exponential similarity function is chosen for its simplicity and suitability for image and feature spaces [16] (here, human faces). It has a fast decreasing rate due to the increase of face distance. Hence, only close-by faces are connected and thus share name labels. This partially satisfies our desire. Because of their variations in expressions, occlusions, and unmoddeled lighting changes (e.g. specularities), not only close-by faces refer to the same person. To moderate the decreasing rate and balance the nearby face connections against the far away face connections, we incorporate the metric σ into this similarity function as follows:

$$sim_{f_i f_j} = exp(-\frac{d_{f_i f_j}^2}{\sigma^2}) = exp(-\frac{\sum_{k=1}^{D}(f_i^k - f_j^k)^2}{\sigma^2}) \quad (5)$$

σ needs to be optimized so that faces in the same cluster should receive higher similarity values than faces in different clusters.

We find σ with a heuristic. First, we find a minimum spanning tree MST over all faces given the Euclidean distances $d_{f_i f_j}$ between faces. Then, we search through the MST for the edge with shortest distance that connects two faces with different labels and get the length d^0 of this edge. We assume d^0 as the minimum distance between face clusters. Following the 3σ rule of the Normal distribution, we set $\sigma = \frac{d^0}{3}$. By using the newly found σ value in estimating the faces similarity, we hope that the similarity of faces within one name class is high and of faces between different name classes is low. This enforces the strong label sharing between faces within one cluster and the weak label sharing between faces from different clusters.

The MST can be found with Kruskal's Algorithm [17]. In our case, the algorithm starts with a graph $G'' =< V'', E'' >$ where the set of vertices V'' contains all faces and the set of edges E'' contains all connections between all faces with weights equal to the face distances. The algorithm returns the minimum spanning tree MST extracted from G''.

Broadcast	P (%)	R (%)	F1 (%)
BBC 22-Jun-2008	100	83.33	90.91
BBC 27-Jun-2008	94.44	94.44	94.44

Table 1. Anchor face track detection performance.

Broadcast	NLF	Percentage
BBC 22-Jun-2008	49	11%
BBC 27-Jun-2008	65	10%

Table 2. NLF: Number of manually labeled faces.

6. EXPERIMENTS

6.1. Evaluation specifics

The performance of the face labeling process is expressed in terms of $Precision$ (P) versus $Recall$ (R). $Precision$ is the proportion of the correctly recognized faces in all recognized faces. The term $Recall$ is the proportion of faces which are assigned a name after the "refusal to predict" mechanism (see section 4.5). We adopt Everingham et al.' approach [4] to evaluate our face labeling method, where we vary the name assignment threshold.

6.2. Datasets and preprocessing

We perform and compare our experiments on two BBC news broadcasts recorded on 22-Jun-2008 and 27-Jun-2008. Each broadcast lasts approximately 30 minutes, or 60,000 frames.

After the face detection and tracking process, we obtain from the *BBC_22-Jun-2008* broadcast 31,275 faces, forming 129 face tracks. From the *BBC_27-Jun-2008* broadcast, 38,487 faces and 169 face tracks are extracted. To reduce the number of false positives in the face detection, we filter out the detections with a too low confidence value. Finally, to further reduce the processing time, *an average* of 3 representative faces are selected for each track. These kept faces are first selected by their size (the larger, the better), then by their fitting confidence. The exact number of representative faces per face track might vary since there are more or less faces having the largest size and fitting confidence. After reduction, the *BBC_22-Jun-2008* broadcast contains 435 faces and 125 face tracks and the *BBC_27-Jun-2008* broadcast contains 670 faces and 168 face tracks.

For the name detection and clustering, we obtain from *BBC_22-Jun-2008* and *BBC_27-Jun-2008* broadcasts 17 and 32 unique candidate names, respectively.

The *BBC_22-Jun-2008* broadcast contains 36 ground truth anchor face tracks while we detect 30 anchor face tracks. The *BBC_27-Jun-2008* broadcast contains 18 ground truth anchor face tracks while we detect 18 anchor face tracks. Table 1 shows the anchor face track detection performance.

6.3. Forming of the training set

After the face filtering and reduction process, from the set of candidate names, each distinct candidate name is used to label

(a) BBC_22-Jun-2008 with solely face similarities.

Recall	50%	70%	90%	100%
NC-NM-NA	90.70	78.33	63.64	59.30
NC-WM-NA	91.11	71.67	60.49	56.98
NC-WM-WA	100.0	100.0	75.00	69.23

(b) BBC_27-Jun-2008 with solely face similarities.

Recall	50%	70%	90%	100%
NC-NM-NA	61.36	46.63	42.50	38.20
NC-WM-NA	80.49	54.10	49.35	43.82
NC-WM-WA	85.00	62.96	47.37	46.34

(c) BBC_22-Jun-2008 with name-face pair similarities.

Recall	50%	70%	90%	100%
WC-NM-NA	93.02	84.75	84.41	82.56
WC-WM-NA	86.05	85.00	84.41	82.56
WC-WM-WA	100.0	84.20	86.36	76.00

(d) BBC_27-Jun-2008 with name-face pair similarities.

Recall	50%	70%	90%	100%
WC-NM-NA	72.09	63.79	46.51	43.48
WC-WM-NA	72.34	65.08	55.00	51.09
WC-WM-WA	72.73	57.14	43.24	36.36

Table 3. Quantitative precision results at different recall levels on the news broadcasts BBC_22-Jun-2008 and BBC_27-Jun-2008.

maximum 6 faces if available (the face selection is random). Note that many names are mentioned in the transcript but do not have corresponding faces in the video frames, so they are assigned to null face. The set of labeled faces will be used as the training set for the name label propagation process. Table 2 shows the number of manual labeled faces per news broadcast.

6.4. Label propagation

We set up the label propagation experiments in three settings: (1) without metric learning for face classification (i.e., using eq. 4) and without anchor detection (denoted as *NM-NA*); (2) with metric learning and without anchor detection (*WM-NA*); and (3) with both metric learning and anchor detection (*WM-WA*). Figure 1 shows precision/recall curves on the news broadcasts *BBC_22-Jun-2008* and *BBC_27-Jun-2008* with name labeling relying solely on the face similarities (*NC-*) and using the name-face pair similarities (*WC-*). And the quantitative precision results at different recall levels are shown in table 3. We can see, the combination of metric learning and anchor detection might be helpful in boosting name-face alignment performance in video. The results of the two episodes show that the label propagation method that uses the name-face pair similarities yields better performance. The incorporation of the named anchors in some cases can not outperform the other methods since it introduces many training examples of some anchor name classes and these anchor name classes attract more faces than other classes do. Moreover, there are a few faces where the system is very confident in assigning names to them, yet that are labeled incorrectly. This might due to the face similarity computation. Hence, this again confirms the difficulty of comparing faces as mentioned in section 5.

We can compare our best results obtained at 100% recall with the precision results of a Support Vector Machine classifier (optimized with RBF kernel) trained with the same labeled faces used in the other experiments. The faces are described with the same features as the ones used in the other experiments. We obtain for the *BBC_22-Jun-2008* broadcast a precision of 55.81%, where our proposed method obtained a precision of 82.56%, and for the *BBC_27-Jun-2008* broadcast a precision of 26.09%, where our proposed method obtained 51.09% precision. A recall level of 100% here again means that after reduction of the number of faces (see table 1), all test faces receive a label.

7. CONCLUSION

We designed, implemented and evaluated a method for naming faces in video broadcasts that learns from labeled and unlabeled examples using iterative label propagation in a graph of connected faces or connected name-face pairs. We used very few labeled data points (about 10-11 % of the detected faces). The label propagation algorithm yields better results than a Support Vector Machine classifier trained on the same labeled data. Our results could be improved by using a similarity metric for comparing faces that uses the distribution of face similarities in a video broadcast.

In our future work we will attempt to further reduce the number of labeled examples, by, for instance, using unsupervised name and faces alignment techniques [18], and keeping the alignments that were recognized with high confidence as the labeled seeds for the propagation algorithm.

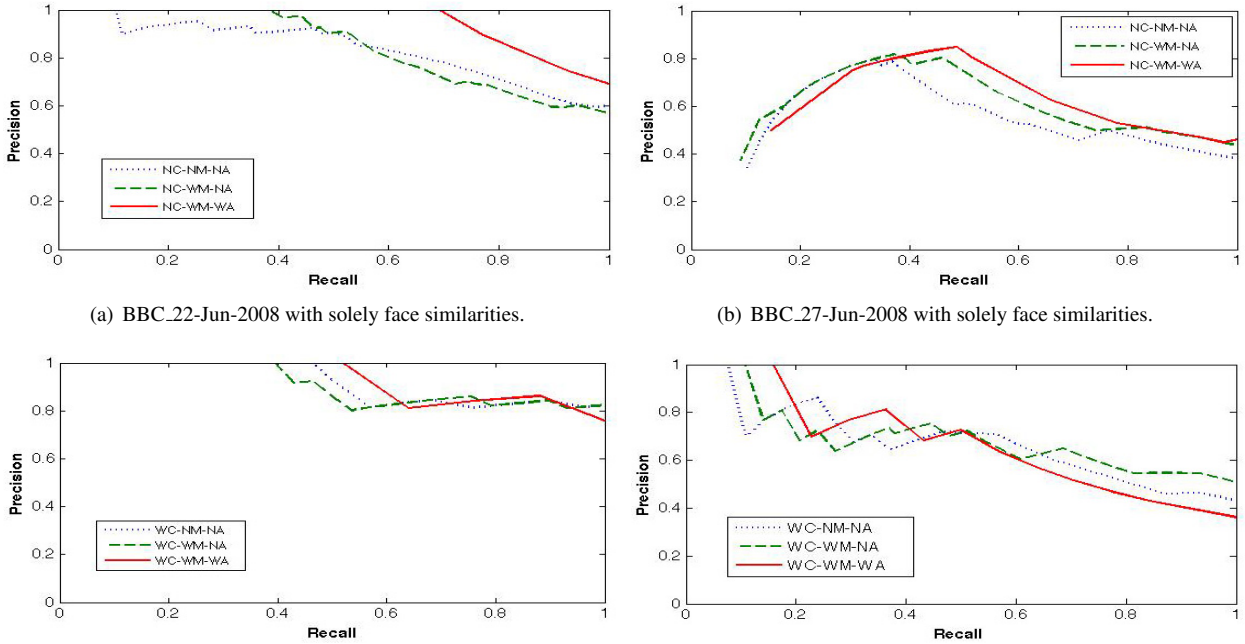
8. REFERENCES

[1] J. Yang and A. G. Hauptmann, "Naming every individual in news video monologues," in *Proceedings of the ACM Multimedia 2004*, October 2004, pp. 580–587.

[2] J. Yang, R. Yan, and A. G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video," in *MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia*. 2005, pp. 31–40, ACM.

[3] S. Satoh, Y. Nakamura, and T. Kanade, "Name-it: Naming and detecting faces in news videos," *IEEE Mutimedia*, vol. 1, pp. 22–35, 1999.

[4] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy – Automatic naming of characters in tv video," in *Proceedings of the British Machine Vision Conference*, 2006.



(a) BBC_22-Jun-2008 with solely face similarities.

(b) BBC_27-Jun-2008 with solely face similarities.

(c) BBC_22-Jun-2008 with name-face pair similarities.

(d) BBC_27-Jun-2008 with name-face pair similarities.

Fig. 1. Precision/recall curves for the news broadcasts BBC_22-Jun-2008 and BBC_27-Jun-2008.

[5] J. Sivic, M. Everingham, and A. Zisserman, ""Who are you?" – Learning person specific classifiers from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[6] D. Ozkan and P. Duygulu, "A graph based approach for naming faces in news photos," in *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2006, pp. 1477–1482, IEEE Computer Society.

[7] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[8] V. Castelli and T. Cover, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *Information Theory, IEEE Transactions on*, vol. 42, no. 6, pp. 2102–2117, 1996.

[9] T. Zhang and F. J. Oles, "A probability analysis on the value of unlabeled data for classification problems," in *Proc. 17th International Conf. on Machine Learning*, 2000, pp. 1191–1198.

[10] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., 2002.

[11] M. Szummer and T. Jaakkola, "Partially labeled classification with Markov random walks," in *Advances in Neural Information Processing Systems*. 2002, pp. 945–952, MIT Press.

[12] L. D'Anna, G. Marrazzo, G. Percannella, C. Sansone, and M. Vento, "A multi-stage approach for anchor shot detection," in *SSPR/SPR*, 2006, pp. 773–782.

[13] P. Viola and M. Jones, "Robust realtime object detection vector quantization," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[14] M. Desmet, R. Fransens, and L. Van Gool, "A generalized em approach for 3D model based face recognition under occlusions," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1423–1430.

[15] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. June 1994, pp. 593–600, IEEE Computer Society Press.

[16] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[17] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[18] P. T. Pham, M.-F Moens, and T. Tuytelaars, "Cross media alignment of names and faces," *IEEE Transactions on Mutimedia*, vol. 12, no. 1, pp. 13–27, 2010.